Evaluation
Experimental protocols, datasets, metrics
Web Search
What makes a good search engine?

- **Efficiency**: It replies to user queries without noticeable delays.
  - 1 sec is the “limit for users feeling that they are freely navigating the command space without having to unduly wait for the computer”

- **Effectiveness**: It replies to user queries with relevant answers.
  - This depends on the interpretation of the user query and the stored information.
### Efficiency metrics

<table>
<thead>
<tr>
<th>Metric name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elapsed indexing time</td>
<td>Measures the amount of time necessary to build a document index on a particular system.</td>
</tr>
<tr>
<td>Indexing processor time</td>
<td>Measures the CPU seconds used in building a document index. This is similar to elapsed time, but does not count time waiting for I/O or speed gains from parallelism.</td>
</tr>
<tr>
<td>Query throughput</td>
<td>Number of queries processed per second.</td>
</tr>
<tr>
<td>Query latency</td>
<td>The amount of time a user must wait after issuing a query before receiving a response, measured in milliseconds. This can be measured using the mean, but is often more instructive when used with the median or a percentile bound.</td>
</tr>
<tr>
<td>Indexing temporary space</td>
<td>Amount of temporary disk space used while creating an index.</td>
</tr>
<tr>
<td>Index size</td>
<td>Amount of storage necessary to store the index files.</td>
</tr>
</tbody>
</table>
What makes a good search engine?

• **Efficiency**: It replies to user queries without noticeable delays.
  - 1 sec is the “limit for users feeling that they are freely navigating the command space without having to unduly wait for the computer”

• **Effectiveness**: It replies to user queries with relevant answers.
  - This depends on the interpretation of the user query and the stored information.
Essential aspects of a sound evaluation

• Experimental protocol
  • Is the task/problem clear? Is it a standard task?
  • Detailed description of the experimental setup:
    • identify all steps of the experiments.

• Reference dataset
  • Use a well known dataset if possible.
    • If not, how was the data obtained?
  • Clear separation between training and test set.

• Evaluation metrics
  • Prefer the commonly used metrics by the community.
  • Check which statistical test is most adequate.
Experimental setups

• There are experimental setups made available by different organizations:

  • TREC: http://trec.nist.gov/tracks.html
  • CLEF: http://clef2017.clef-initiative.eu/
  • SemEVAL: http://alt.qcri.org/semeval2017/
  • Visual recognition: http://image-net.org/challenges/LSVRC/

• These experimental setups define a protocol, a dataset (documents and relevance judgments) and suggest a set of metrics to evaluate performance.
What is a standard task?

• Experimental setups are designed to develop a search engine to address a specific task.
  • Retrieval by keyword
  • Retrieval by example
  • Ranking annotations
  • Interactive retrieval
  • Search query categorization
  • Real-time summarization

• Datasets exist for all the above tasks.
Examples of standard tasks in IR

• For example, TRECVID tasks include:
  • Video shot-detection
  • Video news story segmentation
  • High-level feature task (concept detection)
  • Automatic and semi-automatic video search
  • Exploratory analysis (unsupervised)

• Other forums exist with different tasks:
  • TREC: Blog search, opinion leader, patent search, Web search, document categorization...
  • CLEF: Plagiarism detection, expert search, wikipedia mining, multimodal image tagging, medical image search...
  • Others: Japanese, Russian, Spanish, etc...
A retrieval evaluation setup
Essential aspects of a sound evaluation

- Experimental protocol
  - Is the task/problem clear? Is it a standard task?
  - Detailed description of the experimental setup:
    - identify all steps of the experiments.

- Reference dataset
  - Use a well known dataset if possible.
    - If not, how was the data obtained?
  - Clear separation between training and test set.

- Evaluation metrics
  - Prefer the commonly used metrics by the community.
  - Check which statistical test is most adequate.
Reference datasets

• A reference dataset is made of:
  • a collection of documents
  • a set of training queries
  • a set of test queries
  • the relevance judgments of the pairs query-document.

• Reference datasets are as important as metrics for evaluating the proposed method.
  • Many different datasets exist for standard tasks.
  • Reference datasets set the difficulty level of the task.
  • Allow a fair comparison across different methods.
Ground-truth (relevance judgments)

- Ground-truth tells the scientist how the method must behave.

- The ultimate goal is to devise a method that produces exactly the same output as the ground-truth.

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>True positive</td>
<td>False positive</td>
</tr>
<tr>
<td>False</td>
<td>False negative</td>
<td>True negative</td>
</tr>
</tbody>
</table>

Type I error

Type II error
Annotate these pictures with keywords:
Relevance judgments

People
Nepal
Mother
Baby
Colorful dress
Fence

Sunset
Horizon
Couds
Orange
Desert

Flowers
Yellow
Nature

Beach
Sea
Palm tree
White-sand
Clear sky
Relevance judgments

• Judgments can be obtained by **experts** or by **crowdsourcing**
  • Human relevance judgments can be incorrect and inconsistent

• How do we measure the quality of human judgments?

\[
kappa = \frac{p(A) - p(E)}{1 - p(E)}
\]

- \(p(A)\) -> proportion of times humans agreed
- \(p(E)\) -> probability of agreeing by chance

• Values above 0.8 are considered good
• Values between 0.67 and 0.8 are considered fair
• Values below 0.67 are considered dubious
Essential aspects of a sound evaluation

• Experimental protocol
  • Is the task/problem clear? Is it a standard task?
  • Detailed description of the experimental setup:
    • identify all steps of the experiments.

• Reference dataset
  • Use a well known dataset if possible.
    • If not, how was the data obtained?
  • Clear separation between training and test set.

• Evaluation metrics
  • Prefer the commonly used metrics by the community.
  • Check which statistical test is most adequate.
Evaluation metrics

• Complete relevance judgments
  • Ranked relevance judgments
  • Binary relevance judgments

• Incomplete relevance judgments (Web scale eval.)
  • Binary relevance judgments
  • Multi-level relevance judgments
Ranked relevance evaluation metrics

• Spearman’s rank correlation:

\[ r = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} \]

• Example:

<table>
<thead>
<tr>
<th>1</th>
<th>1</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

\[ r = 1 - \frac{6((1 - 1)^2 + (2 - 3)^2 + (3 - 4)^2 + (4 - 2)^2)}{4(4^2 - 1)} \]

• Another popular rank correlation metric is the Kendall-Tau.
Binary relevance judgments

\[
\text{Accuracy} = \frac{\text{truePos} + \text{trueNeg}}{\text{truePos} + \text{falsePos} + \text{trueNeg} + \text{falseNeg}}
\]

\[
\text{Precision} = \frac{\text{truePos}}{\text{truePos} + \text{falsePos}}
\]

\[
\text{Recall} = \frac{\text{truePos}}{\text{truePos} + \text{falseNeg}}
\]

\[
F_1 = \frac{2}{\frac{1}{P} + \frac{1}{R}}
\]

Ground-truth

<table>
<thead>
<tr>
<th>Method</th>
<th>True</th>
<th>False</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>True positive</td>
<td>False positive</td>
</tr>
<tr>
<td>False</td>
<td>False negative</td>
<td>True negative</td>
</tr>
</tbody>
</table>

Em PT: exatidão, precisão e abrangência.
Precision-recall graphs for ranked results

- Improved precision
- Improved F-measure
- Improved recall

Precision vs. Recall graph with systems A, B, and C.
Interpolated precision-recall graphs
Average Precision

- Web systems favor high-precision methods (P@20)

- Other more robust metric is AP:

\[
AP = \frac{1}{\text{#relevant}} \cdot \sum_{k \in \{\text{set of positions of the relevant docs}\}} p@k
\]

\[
AP = \frac{1}{4} \cdot \left( \frac{1}{2} + \frac{2}{4} + \frac{3}{6} \right) = 0.375
\]
Average Precision

- Average precision is the area under the P-R curve

\[
AP = \frac{1}{\#\text{relevant}} \cdot \sum_{k \in \{\text{set of positions of the relevant docs}\}} p@k
\]
Mean Average Precision (MAP)

- MAP evaluates the system for a given range of queries.
- It summarizes the global system performance in one single value.
- It is the mean of the average precision of a set of $n$ queries:

\[
MAP = \frac{AP(q_1) + AP(q_2) + AP(q_3) + \ldots + AP(q_n)}{n}
\]
Web scale evaluation

• It is impossible to know all relevant documents.
  • It is too expensive or time-consuming.

• **DCG**, **BPref** and **Inferred AP** are three measures to evaluate a system with incomplete ground-truth.

• These metrics use the concept of pooled results

---

Results pooling

• This technique is used when the dataset is too large to be completely examined.

• Considering the results of 10 systems:
  • Examine the top 100 results of each system
  • Label all documents according to its relevance
  • Use the labeled results as ground-truth to evaluate all systems.

• **Drawback: can’t compute recall, AP and MAP**
DCG: Incomplete multi-level relevance

• Useful when some documents are more relevant than others.

• Documents need to have ground-truth with different levels of relevance.

• A common metric is the Discounted Cumulative Gain:

\[
DCG_m = \sum_{i=1}^{m} \frac{2^{rel_i} - 1}{\log_2(1 + i)} \quad rel_i = \{0,1,2,3,...\} \quad nDCG_m = \frac{DCG_m}{bestDCG_m}
\]

BPref: Incomplete binary relevance

• When only incomplete binary relevance judgments are available B Pref is a popular metric:

\[
BPREF = \frac{1}{R} \sum_{d_r} \left(1 - \frac{N_{d_r}}{R}\right)
\]

• where \( R \) is the total number of relevant documents in a given query

• \( d_r \) is a relevant document

• \( N_{d_r} \) is the number of non-relevant documents ranked higher than \( d_r \)
Diversity and novelty

• Diversity and novelty are difficult to evaluate.

• There are no defacto method to measure it.

• The goal is to measure **how diverse and novel is the information** contained in the retrieved documents.
  • Assessment focus is not at the level of the documents.
Nuggets or information facts

• A **nugget** is an information fact
  • **Documents** contain many nuggets.
  • The same **nugget** can be present in many different documents.

• The goal is to retrieve a ranked list with many different nuggets at the top of the list

• Repeated nuggets will have a decreasing importance
The $\alpha$-nDCG metric for diversity and novelty

- The relevance of a document is determined by its nuggets

$$\sum_{j=1}^{m} N(d_i, n_j).$$

and by the nuggets that occurred previously in the ranked results

$$r_{j,k-1} = \sum_{i=1}^{k-1} N(d_i, n_j),$$

- A popular metric is the $\alpha$-nDCG, where each document at position $k$ is judged by its nuggets

$$G[k] = \sum_{j=1}^{m} N(d_k, n_j) \alpha^{r_{j,k-1}}, \quad \alpha = 0.5$$
Example

• Top results for query “Norwegian Cruise Lines”

<table>
<thead>
<tr>
<th>Document Title</th>
<th>85.1</th>
<th>85.2</th>
<th>85.3</th>
<th>85.4</th>
<th>85.5</th>
<th>85.6</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Carnival Re-Enters Norway Bidding</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>b. NORWEGIAN CRUISE LINE SAYS...</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>c. Carnival, Star Increase NCL Stake</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>d. Carnival, Star Solidify Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>e. HOUSTON CRUISE INDUSTRY GETS...</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>f. TRAVELERS WIN IN CRUISE...</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>g. ARMCHAIR QUARTERBACKS NEED...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>h. EUROPE, CHRISTMAS ON SALE</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>i. TRAVEL DEALS AND DISCOUNTS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>j. HAVE IT YOUR WAY ON THIS SHIP</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
</tbody>
</table>

\[
r_{j,k-1} = \sum_{i=1}^{k-1} N(d_i, n_j),
\]

\[
G[k] = \sum_{j=1}^{m} N(d_k, n_j) \alpha^{r_{j,k-1}},
\]

• The relevance of each document is: \( G = \langle 2, \frac{1}{2}, \frac{1}{4}, 0, 2, \frac{1}{2}, 1, \frac{1}{4}, \ldots \rangle. \)

• What would be the ideal ordering?

\( a-c-g-b-f-c-h-i-j-d \) \( G' = \langle 2, 2, 1, \frac{1}{2}, \frac{1}{2}, \frac{1}{4}, \frac{1}{4}, \ldots \rangle. \)
System quality and user utility

• The discussed evaluation procedures only measure the system performance on a given task
  • It can overfit
  • It might be distant from what users expect

• Only real users actually assess the system
  • How expressive is its query language?
  • How large is its collection?
  • How effective are the results?

• A/B testing
  • Make small variation on the system and direct a proportion of users to that system
  • Evaluate frequency with which users click on top results
Qualitative discussion

• Relevance depends on:
  • Task objective
  • User knowledge
  • Time

• Not all people “see” the same
  • Binary relevance judgments
  • Multi-level relevance judgments
  • Ranked relevance judgments
  • Incomplete relevance judgments

The notion of relevance is a subjective concept

There is no relation between AP and user satisfaction
Summary

• Metrics for complete relevance judgments
  • Binary: Precision, Recall, F-measure, Average Precision, Mean AP
  • Ranked: Spearman, Kendall-tau

• Metrics for incomplete relevance judgments
  • Binary: Bpref, InfMAP
  • Multi-valued: Normalized DCG

• Evaluation collections / resources
  • See TRECVID and ImageCLEF for multimedia datasets.
  • See TREC and CLEF forums for Web and large-scale datasets
    • User search interaction, Geographic IR, Expert finding, Blog search, Plagiarism,…
  • Use trec_eval application to evaluate your system